Multiple Sequence Alignment

Lessons from the School of Hard Knocks

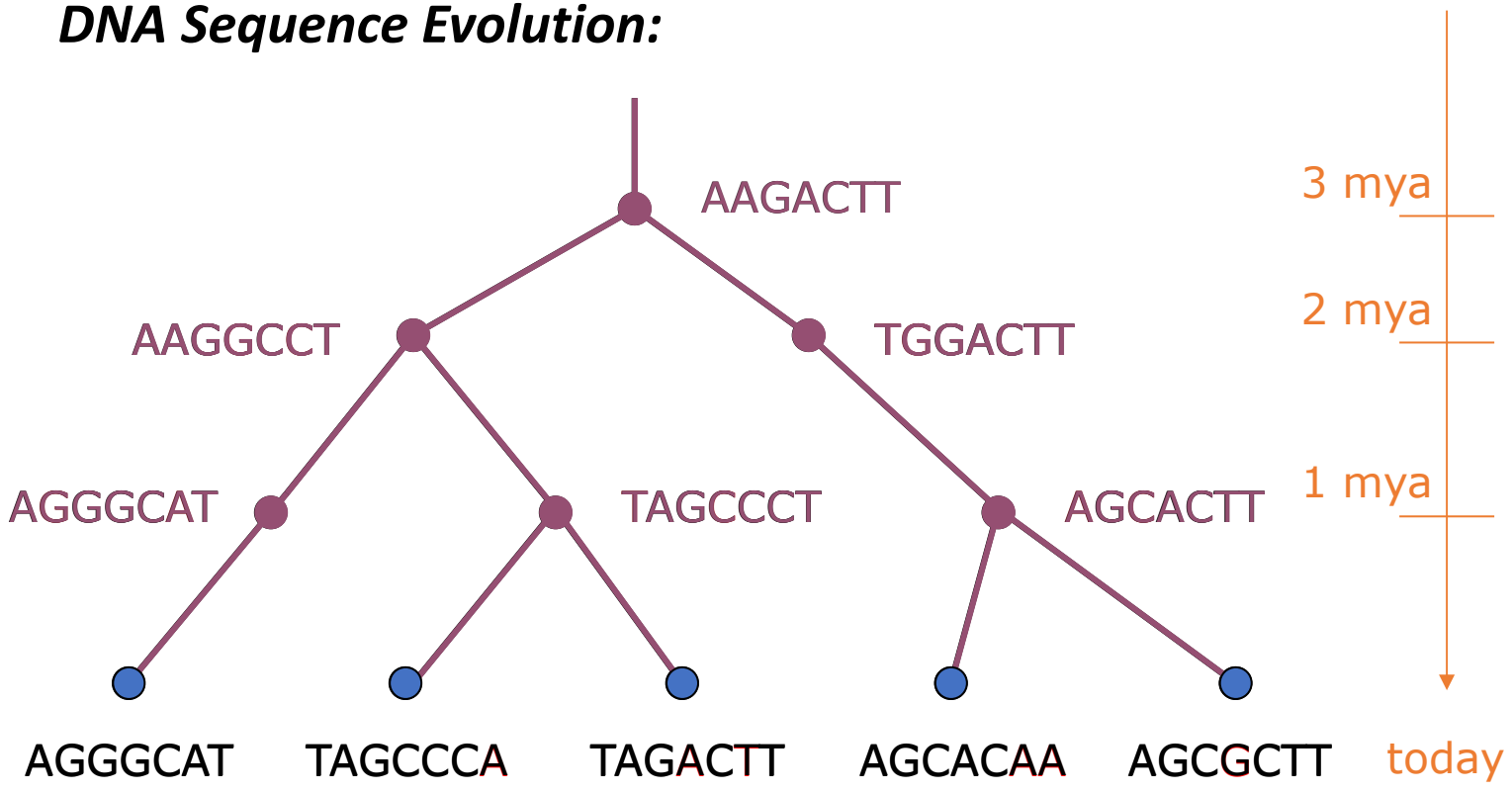
Michael Nute

STAMPS 2022

July 29, 2022

Brief Intro to Molecular Phylogenetics

DNA Sequence Evolution:



Notes:

- Only character mutation shown. Other operations are possible:
  - Insertion and deletion
  - Duplication
  - Transposition
  - Etc…
- Observed data are the "extant" sequences (at the bottom of the tree).

Two Separate-but-Related Problems:

1. Identify which groups of characters share a common ancestor. (Multiple Sequence Alignment)
2. Identify topological structure of the evolutionary history. (Phylogeny Estimation)

*mya = million years ago

*Slide Credit: Prof. Tandy Warnow

Multiple Sequence Alignment: Definition & Goal

Input: Sequences from different organisms (or different loci) that evolved from a common ancestor.

Goal: Align sequences so that all sets of positions having a common ancestor are grouped together.

- *Not the same as aligning short sequences (or reads) to a reference ("mapping").*
- *Not the same as **genome** alignment*
- *Typically done before creating a phylogenetic tree…*

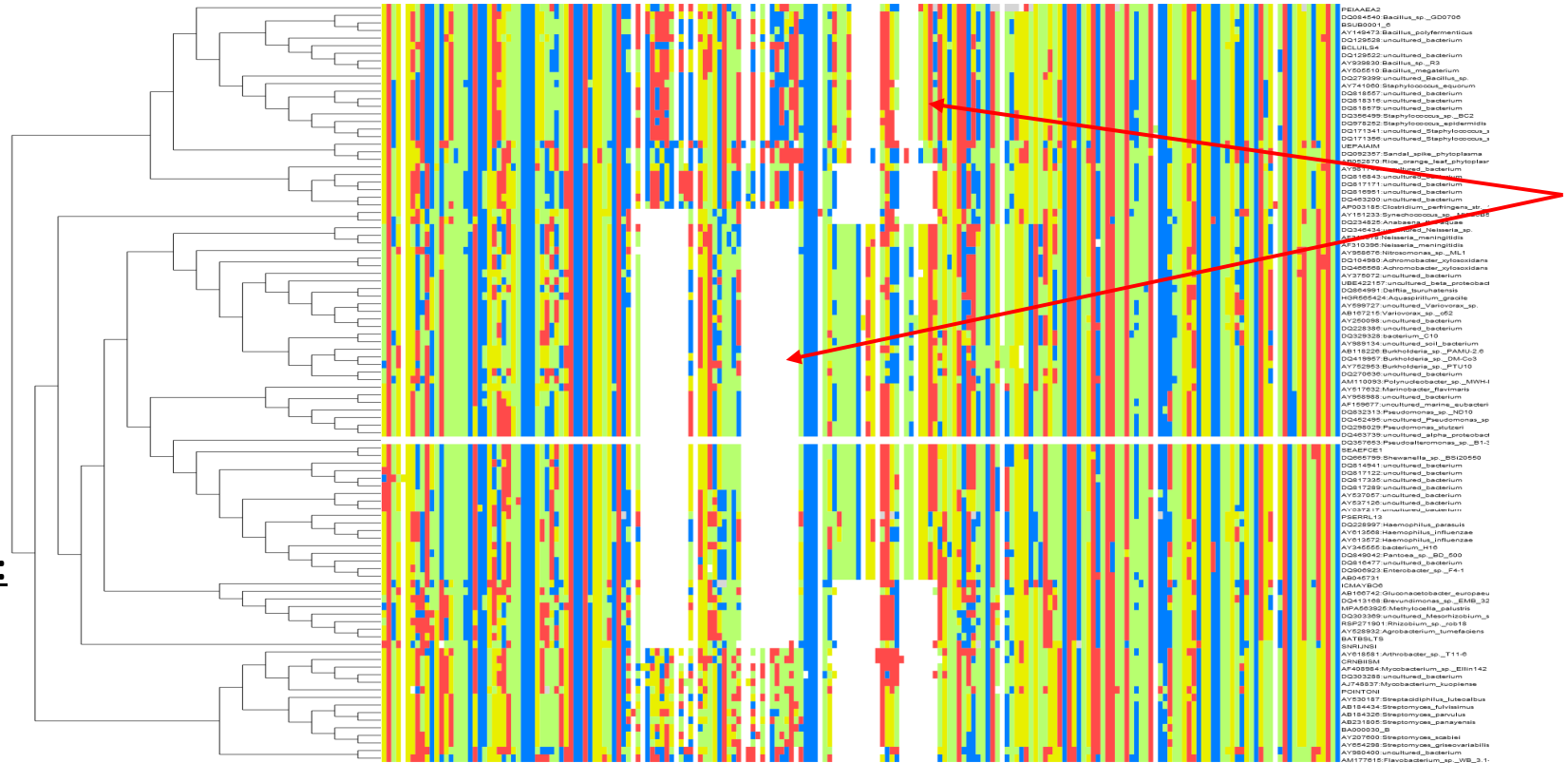
Tools:

- MAFFT
- Muscle
- PASTA
- ClustalW
- DiAlign
- BAli-Phy
- PRANK
- T-COFFEE
- *…et cetera*

Garbage Alignment → ***Garbage Tree***

Multiple Sequence Alignment (Single-Gene)

- Goal: Align the sequences so that each set of “aligned” characters evolved from a common ancestral character.
  - Typically only allow mutation and insertion/deletion
    - (i.e. aligned sequences must have same characters, in same order, as originals).
  - Related to problem of phylogeny estimation.



White indicates a “gap” in the alignment, a.k.a. an insertion or deletion (indel)

Legend:

Data Properties Affecting Multiple Sequence Alignment

- Avg Sequence Similarity (rate of evolution)
- # of Sequences
- Presence of highly conserved regions
- Sequence Length heterogeneity
- Gap length/frequency
- Sequence fragmentation (*not the same as heterogeneity*)
- Avg Sequence Length

MSA Algorithms & Software (Partial List)

Tool	Use Case	Comments
MAFFT (Katoh et al., 2002)	Single Gene MSA (small N)	• Uses patterns of insertion/deletion to find optimal alignment • Generally pretty accurate in most conditions.
MUSCLE (Edgar, 2004)		• Progressive alignment. Suitable for relatively high overall sequence similarity.
CLUSTAL (Sievers et al., 2011)		• Ideal for protein alignments with structurally important sites.
PASTA (Mirarab et al., 2015)	Single Gene MSA (large N)	• Divide-and-conquer algorithm. Ideal for scaling alignment to large number of sequences (>1000)
HMMER (Eddy, 1998)	*Query sequence alignment to reference ("mapping")*	*• Represents reference alignment as HMM. Query sequence alignment performed using standard HMM algorithms.*

I tend to tell people:

- *Just run MAFFT for anything less than 500 sequences or so*
- *Muscle is fine too if the divergence is low…*
- *…or ClustalW for AA sequences with important structural sites*
- *Use PASTA for over 1k sequences or if avg. %-identity is very low (high rate of evolution).*

This simplistic advice is a STAMPS 2022 exclusive…

Large N, Low %ANI ⟶ Very Hard Alignment

It is far easier than widely appreciated to get an alignment with ≈0% accuracy.

BE CAREFUL OUT THERE!

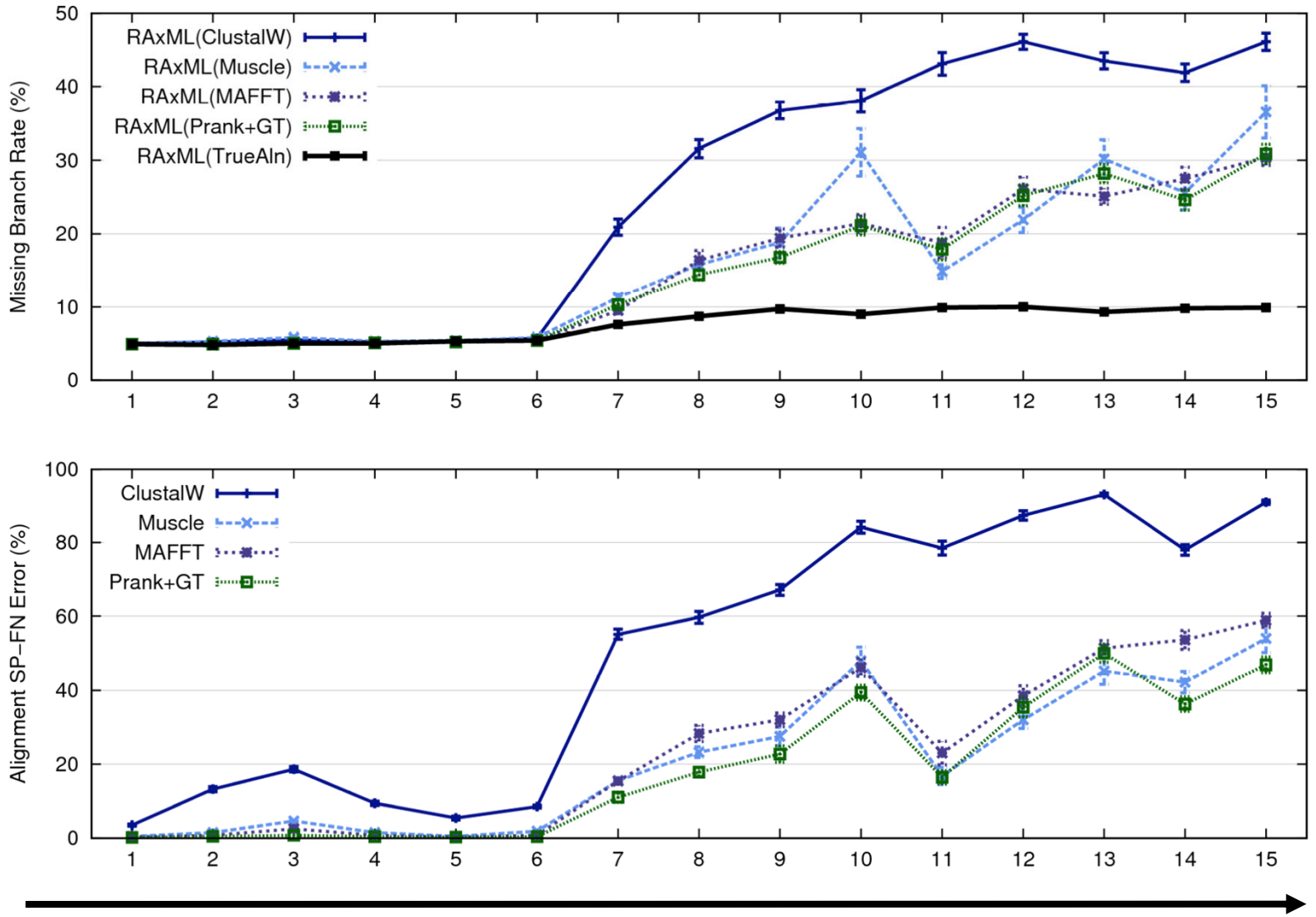
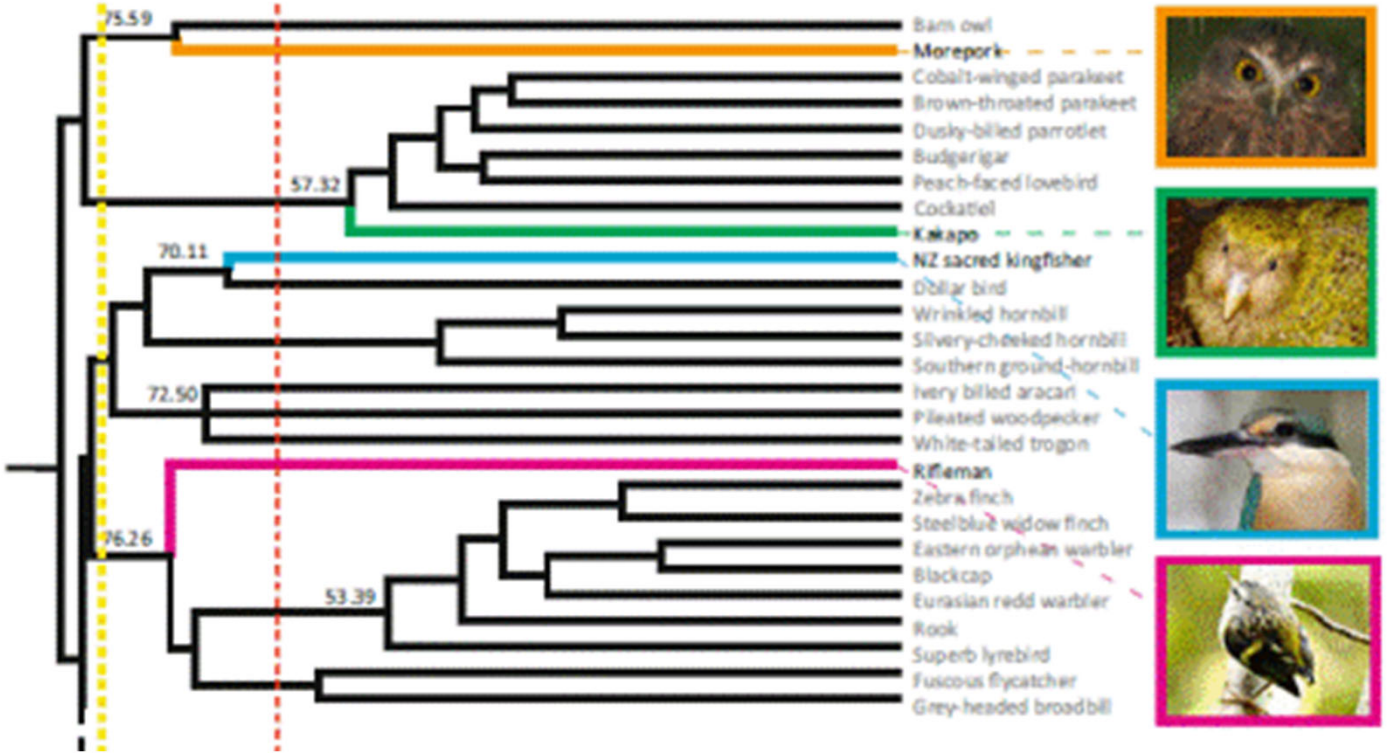


Image Credit: Tandy Warnow.

What Happens If the Alignment is Garbage?

A poor alignment will give a tree with *very deep* branches

- I.e. long branches above leaves.
  - A “Star”-like tree
- “Nothing in this tree is systematically related to anything else in any significant capacity.”
  - *Could be because relationships were nuked by bad alignment*
- Of course, star-like trees *can* be real! (e.g. birds)



Failure Modes: Under/Over-Alignment

- Most of the commonly used alignment methods will tend to *over*-align.
- *…better than aligning just the right amount incorrectly!*
- *…but can lead to some weird internal branches…*